Modelización de Sistemas Biológicos

Componente Aleatoria de los Modelos

POLITÉCNICA

Componente Aleatoria de los Modelos

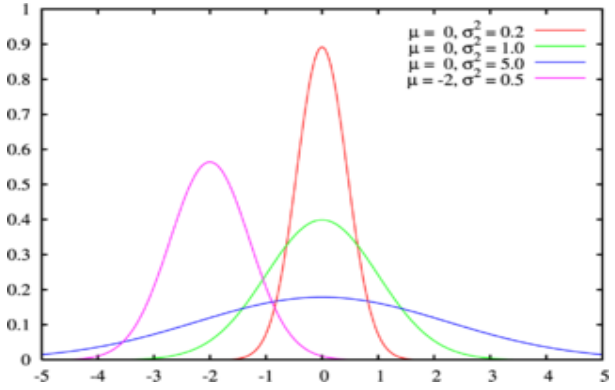
Variables Aleatorias

Variables de Entrada

- Está muy vinculado con el conocimiento que se tiene del fenómeno a estudiar:
  - Asignación de distribuciones basándose en hipótesis: Susceptible de error y no corroborado por datos. Válido si se usan diferentes suposiciones complementarias.
  - Estimación de distribuciones a partir de los datos: Diseño experimental de la recogida de datos o datos disponibles y ajuste de los mismos a distribuciones:
    - Selección de distribución: Continuas o discretas.
    - Ajuste de parámetros.

Distribuciones Continuas

Normal/Gaussiana



N(μ,σ)

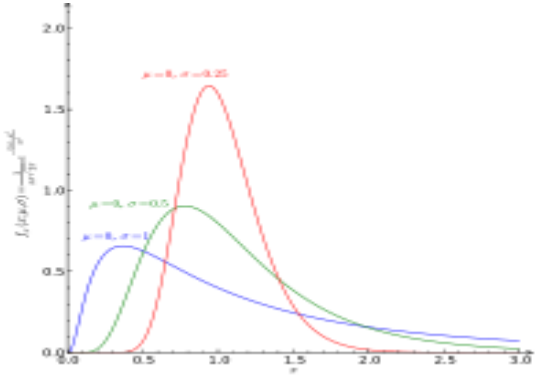
Tests:

- Kolmogorov-Smirnov
- Shapiro-Wilk
- K^2 de D'Agostino (momentos)

Ajuste:

- Estimador de máxima verosimilitud.
- Estimador de Bessel

Log-Normal



ln N(μ,σ)

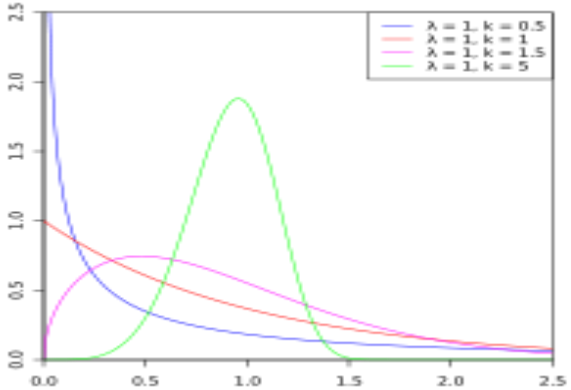
Tests:

- El logaritmo es una variable con distribución normal

Ajuste:

- Estimador de máxima verosimilitud.

Weibull



W(λ,k)

Estimadores:

$$\hat{\lambda} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i^{\ k} - x_N^{\ k}\right)$$

$$\hat{k} = \left(\frac{\sum_{i=1}^{N}\left(\ln(x_i)x_i^{\ k} - \ln(x_N)\,x_N^{\ k}\right)}{\sum_{i=1}^{N}\left(x_i^{\ k} - x_N^{\ k}\right)} - \frac{1}{N}\sum_{i=1}^{N}\ln(x_i)\right)^{-1}$$

Distribuciones Continuas (cont.)

- Otras distribuciones continuas:
  - Exponencial
  - Fisher
  - Gamma
  - Logística
  - Uniforme
  - Wald
  - …

El número de distribuciones y la tipología es muy variada. La aproximación habitual es considerar:

1. Probar por medio de tests las que son más probables (uniforme, normal o lognormal).
2. Si los resultados no son concluyentes, entonces considerar una serie (4-5) de alternativas.
3. Aplicar los estimadores de los parámetros que las caracterizarían.
4. Comparar la calidad del ajuste de una distribución con esos parámetros con la serie de datos.

Distribuciones Discretas

- Diferentes tipos:
  - Binomial.
  - Uniforme discreta.
  - Bernouilli.
  - Poisson.
  - Boltzmann discreta.
  - Hipergeométrica.

En las distribuciones discretas hablamos de función de masa de probabilidad en lugar de densidad de probabilidad.

- El cálculo analítico de estas distribuciones se basa en el sumatorio de series y en combinatoria y no en la integración.
- En la práctica, aunque hay casi la misma cantidad de tipos, terminan ajustándose bien uniformes, binomiales y Poisson.

Procedimiento de Ajuste

- Estimadores de máxima verosimilitud:
  - Se deben considerar las funciones de densidad (masa) de probabilidad: Su fórmula analítica (*pdf:* $\mathcal{L}$).
  - Como las funciones suelen ser "complicadas" de resolver se suele aplicar el logaritmo de dicha función.
  - Y para maximizarla se deriva respecto a cada uno de los parámetros.

Ejemplo (función Gamma):

$$\mathcal{L}(\alpha,\beta|x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

logaritmo

$$\ln(\mathcal{L}(\alpha,\beta|x)) = \alpha\ln(\beta) - \ln\big(\Gamma(\alpha)\big) + (\alpha-1)\ln(x) - \beta x$$

Maximizar (derivada igual a 0)

$$\frac{\partial}{\partial\beta}\ln(\mathcal{L}(\alpha,\beta|x)) = 0 \Rightarrow \alpha\beta^{-1} - x = 0 \Rightarrow \beta = \frac{\alpha}{x}$$

Para *n* muestras aleatorias independientes

$$\hat{\beta} = \frac{\alpha}{\bar{x}}$$

Procedimiento de Ajuste (cont.)

- Método de los Momentos:
  - Método de máxima verosimilitud analíticamente complejo (no vale para todas las distribuciones).
  - Los momentos son fáciles de calcular:
    - Momento de orden *k*: $\mu_k = E(X^k) = \frac{1}{N}\sum_{i=1}^{N} X_i^{\ k}$
    - Momentos centrados: $m_k = E((X-\mu_1)^k) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^m$
  - Procedimiento:
    1. Dada una distribución candidata (hipótesis de partida) que tiene varios parámetros.
    2. Se determina analíticamente los momentos 1 a *n* (siendo *n* el número de parámetros de la distribución).
    3. Se calcula los momentos de la muestra disponible.
    4. Se despejan la *n* ecuaciones con *n* incógnitas que salen.

Procedimiento de Ajuste (cont.)

- Método de los Momentos (cont.):
  - Para muchas de las distribuciones estos momentos ya están calculados, o al menos algunas de las versiones normalizadas de los mismos:
    - Media: $\bar{x} = \mu_1$

    Medida de centricidad.
    - Varianza: $\sigma^2 = m_2 = \mu_2 - {\mu_1}^2$

    Medida de variación sobre la centricidad
    - Asimetría: $g_1 = \frac{m_3}{{m_2}^{3/2}}$

    Diferencias de variación a un lado y a otro de la media (*skewness*).
    - Curtosis: $g_2 = \frac{\mu_4}{\sigma^4} - 3$

    Medida de forma (*kurtosis*).

Procedimiento de Ajuste (cont.)

- Método de los Momentos (ejemplo):
  - Suponemos que la serie de valores:
  - sigue una distribución uniforme entre *a* y *b*.

1,39
1,47
0,97
0,87
1,73
1,23
1,89
1,55
1,66
1,25
1,36
1,34
1,72
1,08
1,35
1,69
0,81
1,40
1,13
1,44
1,35
1,13

$$\mathcal{L}(x) = \begin{cases} \dfrac{1}{b-1} & a \leq x \leq b \\ 0 & otro\ caso \end{cases}$$

$$\mu_1 = E(X) = \int_a^b \frac{x}{b-a} dx = \frac{a+b}{2}$$

$$\hat{a} = \hat{\mu}_1 - \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1{}^2)} = 0.870$$

$$\mu_2 = E(X^2) = \int_a^b \frac{x^2}{b-a} dx = \frac{a^2 + ab + b^2}{3}$$

$$\hat{b} = \hat{\mu}_1 + \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1{}^2)} = 1.840$$

Procedimiento de Ajuste (cont.)

- Cuando las funciones son muy complejas (por lo general combinaciones de funciones) los métodos de ajuste se complican y se tiene que resolver por medio de una técnica de optimización:
  - Optimización numérica: Método de Newton y similares.
  - Optimización heurística: Búsqueda local, algoritmos evolutivos, métodos híbridos, …
- Qué es lo que se optimiza:
  - Algún estadístico que mida la bondad del ajuste.
  - Momentos como la media, varianza, asimetría o curtosis.

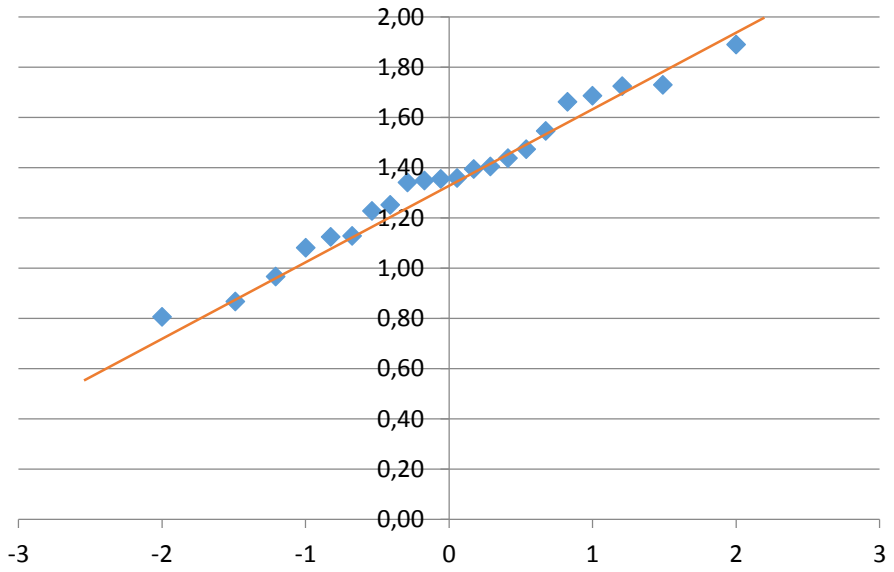
Procedimiento de Ajuste (cont.)

- Bondad de ajuste (*Goodness-of-fit*): Se refiere a una serie de estadísticos (medidas) para determinar cómo de bien se ajusta una distribución parametrizada a unos datos:
  - Métodos visuales: Indicadores visuales de los ajustes.
  - Métodos estadísticos: Medidas objetivas de los valores de ajuste.

- Este tipo de cálculos viene implementado en cualquier paquete estadístico (R, SPSS, SAS, …).

Procedimiento de Ajuste (cont.)

- Bondad de ajuste (cont.): Métodos visuales
  - P-P (probability-probability) Plots: Ajuste a la normalidad.
  - Q-Q (quantil-quantil) Plots: Caso general del anterior.



1. Se ordenan los valores de la muestra y se asignan un valor de ranking (posición): 1, 2, 3, …
2. Se comparan con los valores de los z-scores (inversa de la función de probabilidad acumulada) de una normal.

Time	Rank	Z-score
0,81	1	-2,00042357
0,87	2	-1,48947004
0,97	3	-1,20741405
1,08	4	-0,99820117
1,13	5	-0,82549449
1,13	6	-0,67448975
1,23	7	-0,53751911
1,25	8	-0,40998332
1,34	9	-0,28880936
1,35	10	-0,17174709
1,35	11	-0,05699967
1,36	12	0,05699967
1,39	13	0,17174709
1,40	14	0,28880936
1,44	15	0,40998332
1,47	16	0,53751911
1,55	17	0,67448975
1,66	18	0,82549449
1,69	19	0,99820117
1,72	20	1,20741405
1,73	21	1,48947004
1,89	22	2,00042357

Procedimiento de Ajuste (cont.)

- Bondad de ajuste:
  - Test de la χ^2: Aplicable tanto a problemas discretos como continuos:
    - Se definen una serie de intervalos *N*.
    - Se contabilizan el número de valores de la muestra en cada intervalo *i*: O_i
    - Se contabilizan el número de valores de un muestreo de la distribución candidata para cada intervalo *i*: E_i
    - El paso anterior se puede generalizar si se dispone de una expresión analítica para integrar la función de densidad.
    - Se calcula el valor:

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$$

Intervalos

Variables del modelo

    - Se compara con el valor de una distribución χ^2 con *N-k-1* grados de libertad para un p-valor α (nivel de confianza).

Procedimiento de Ajuste (cont.)

- Bondad de ajuste:
  - Test de la χ^2: Ejemplo
    - Intentamos establecer el grado de ajuste de la muestra con una distribución uniforme y una normal.

	[0,8-1,06)	[0,8-1,32)	[1,06-1,58)	[1,32-1,84)	[1,58-2,1)	
O	3	5	9	4	1	
E U(a,b)	4	5	7	5	1	
E N(μ;σ)	3	5	6	5	3	Chi2
Chi2 U(a,v)	0,25	0	0,57142857	0,2	0	1,02142857
Chi2 N(μ;σ)	0	0	1,5	0,2	1,33333333	3,03333333

La hipótesis de que la distribución se corresponde con una Uniforme (*a,b*) no se verifica con un nivel de confianza del 90% (1,0-α)

Chi2	Chi2 (2 grados)	
0,01	α=0,01	0,02010067
0,05	α=0,05	0,10258659
0,1	α=0,10	0,21072103

O	E U(a,b)	E N(μ;σ)
0,81	0,87963847	0,80937049
0,87	0,94577138	0,95470546
0,97	0,94949072	1,03912496
1,08	1,04858827	1,10975207
1,13	1,14284374	1,13854257
1,13	1,16639693	1,16331845
1,23	1,17216567	1,18617849
1,25	1,30243464	1,20338644
1,34	1,30450331	1,32301122
1,35	1,43177505	1,36468897
1,35	1,48210866	1,36567037
1,36	1,53573428	1,45284203
1,39	1,54167627	1,47305932
1,40	1,54503063	1,52313937
1,44	1,56061049	1,59305711
1,47	1,5676959	1,59664188
1,55	1,60036757	1,62151324
1,66	1,6307297	1,65835823
1,69	1,68939018	1,65903139
1,72	1,8000471	1,88951885
1,73	1,82920994	1,89885546
1,89	1,85617618	2,02884399

Procedimiento de Ajuste (cont.)

- Bondad de ajuste (alternativas):
  - Cálculo empírico de la función de distribución de probabilidad acumulada.
  - Test exacto de Fisher: Valores discretos con pocos valores.
  - Anderson-Darling.
  - Shapiro-Wilk.
  - Kolmogorov-Smirnov.
  - Criterio de Cramér–von Mises.
  - …

Variables de Entrada: Ejemplo

- Volvemos al ejemplo del cálculo de coste (tema de Simulaciones de Monte Carlo):
  - La variable de entrada que debemos ajustar es el número de solicitudes.
  - Para ello disponemos de la siguiente serie de datos (suponemos que recogidos de forma fiable).
  - Suponemos que puede ajustarse a:
    - Normal.
    - Lognormal.
    - Gamma.

Muestra
2076
3068
2329
2237
2575
2234
2372
2336
1555
2139
3184
2312
1917
2283
2574
2992
2591
2773
2234
2563
2069
2948
2801
1852
2248
2741
2218
2715

⋮

Variables de Entrada: Ejemplo (Cont.)

- Utilizaremos R:

R Console

```
> library(MASS)
> fichero = read.csv("D:/Teaching/CpI/Monte Carlo Simulations/Data/coste.csv")
> datos=c(fichero)
> nor<-fitdistr(datos$Muestra, "normal")
> nor
      mean          sd
  2534.90800    709.10886
 (  44.84798) (  31.71231)
> lnor<-fitdistr(datos$Muestra, "log-normal")
> lnor
    meanlog       sdlog
  7.80044118   0.27280180
 (0.01725350) (0.01220007)
> gam<-fitdistr(datos$Muestra, "gamma")
> gam
      shape           rate
  1.350814e+01   5.328848e-03
 (9.370359e-01) (3.699069e-04)
```

Referencias:
http://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf
http://stat.ethz.ch/R-manual/R-patched/library/MASS/html/00Index.html

Variables de Entrada: Ejemplo (Cont.)

R Console

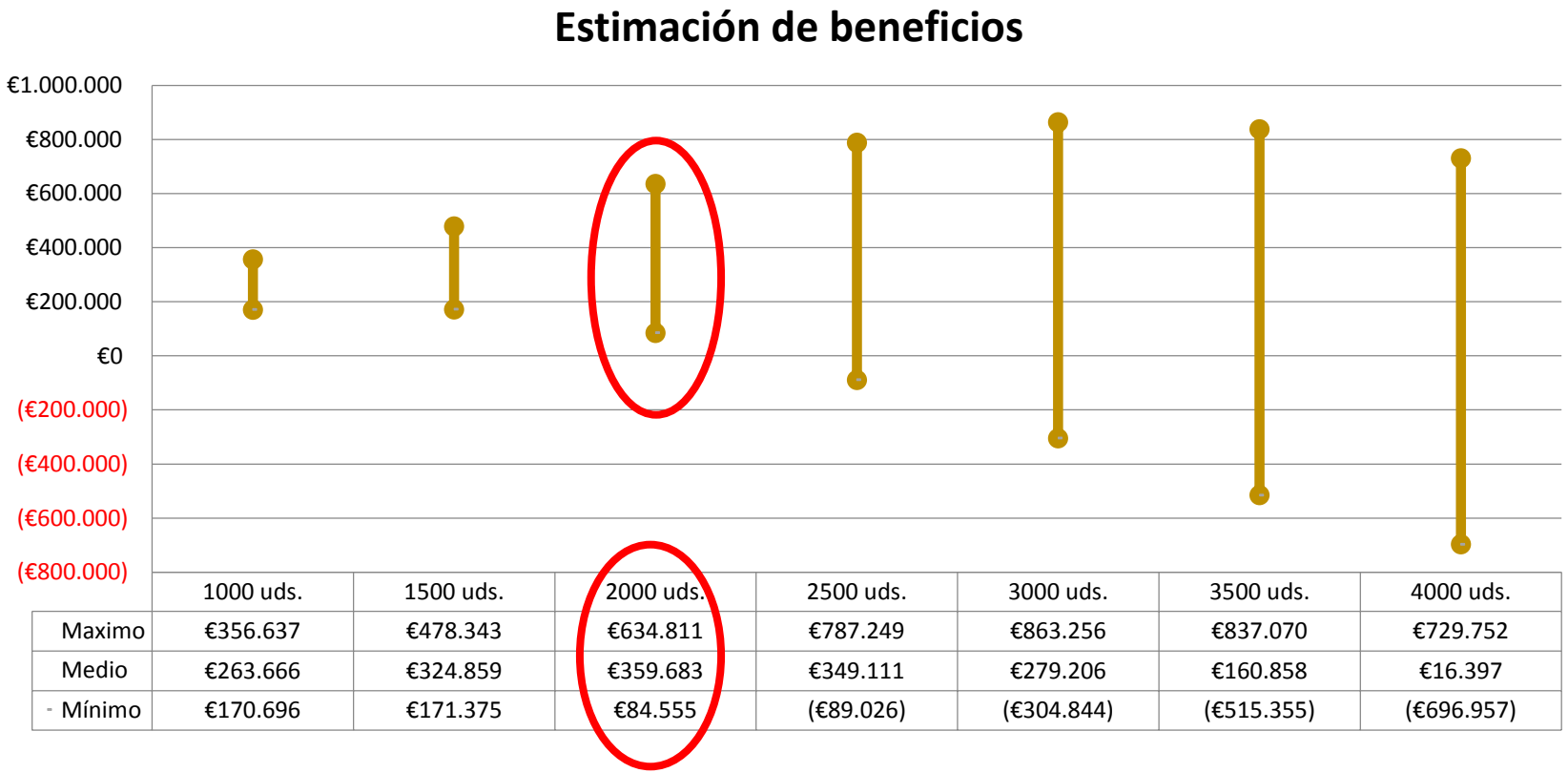
```
> library(vcd)
> intervalos<-c(1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000)
> observaciones<-cut(datos$Muestra, breaks=intervalos)
> obs<-vector()
> for(i in 1:8) obs[i]<- table(observaciones)[[i]]
> obs
[1]  8 50 78 56 31 19  4  4
>
> exp.gam<-vector()
> for(i in 1:8)
+   exp.gam[i]<-(pgamma(1000+(i*500),shape=gam$estimate[1:1],rate=gam$estimate[2:2]) -
+                pgamma( 500+(i*500),shape=gam$estimate[1:1],rate=gam$estimate[2:2])) * 250
> X2.gam<-sum(((obs-exp.gam)^2)/exp.gam)
> 1-pchisq(X2.gam,8-2-1)
[1] 0.1235806
>
> exp.nor<-vector()
> for(i in 1:8)
+   exp.nor[i]<-(pnorm(1000+(i*500),mean=nor$estimate[1:1],sd=nor$estimate[2:2]) -
+                pnorm( 500+(i*500),mean=nor$estimate[1:1],sd=nor$estimate[2:2])) * 250
> X2.nor<-sum(((obs-exp.nor)^2)/exp.nor)
> 1-pchisq(X2.nor,8-2-1)
[1] 5.530155e-06
>
> exp.lnor<-vector()
> for(i in 1:8)
+   exp.lnor[i]<-(plnorm(1000+(i*500),mean=lnor$estimate[1:1],sd=lnor$estimate[2:2]) -
+                 plnorm( 500+(i*500),mean=lnor$estimate[1:1],sd=lnor$estimate[2:2])) * 250
> X2.lnor<-sum(((obs-exp.lnor)^2)/exp.lnor)
> 1-pchisq(X2.lnor,8-2-1)
[1] 0.5018079
```

Variables de Entrada: Ejemplo (Cont.)

- Se puede asumir que las variables de entrada (número de peticiones):
  - Sigue una distribución Normal.
  - Con media μ=2534,9
  - Con desviación estándar σ=709,1

Variables de Entrada: Ejemplo (Cont.)

- Simulación de 1000 escenarios, presentación de los valores medios ± 2σ



	1000 uds.	1500 uds.	2000 uds.	2500 uds.	3000 uds.	3500 uds.	4000 uds.
Maximo	€356.637	€478.343	€634.811	€787.249	€863.256	€837.070	€729.752
Medio	€263.666	€324.859	€359.683	€349.111	€279.206	€160.858	€16.397
Mínimo	€170.696	€171.375	€84.555	(€89.026)	(€304.844)	(€515.355)	(€696.957)

Componente Aleatoria de los Modelos

Generación de Números Aleatorios

Generadores de Números Aleatorios

Nos referimos como RNG (*Random Number Generators*)
Objetivo ideal:

- Uniformemente distribuido
- No correlado
- No se producen ciclos
- Independiente de máquina
- Reproducible
- Satisfacen tests de aleatoriedad
- Cambio en la "semilla" afecta a los valores
- Divisible en secuencias
- De ejecución rápida
- Consumo de memoria razonable

En la práctica: Un ordenador no es aleatorio, de forma que la generación computacional de números es sólo ***pseudoaleatoria***:

- Aritmética con precisión finita ➔ Se producen ciclos
- Reproducible ➔ Existen correlaciones entre semilla y valores.
- Habitualmente se prima la velocidad frente a la calidad de los resultados (al menos en las implementaciones de sistema).

Generadores de Números Aleatorios

- RNG de congruencia lineal (LCG):
  - $X_i = (a \times X_{i-1} + c)\ mod\ M$
  - La semilla fija X_0
  - Produce ciclos con frecuencia (en el mejor de los casos) *M*.
  - *M* está relacionado con la longitud en bits del formato de número aleatorio.
  - Recomendable >42 bits y M no ser una potencia de 2 (evita la correlación de los bits menos significativos).
  - Para producir números en coma flotante, se divide entre el valor de *M*: $X \sim \mathcal{U}(0,1)$

Generadores de Números Aleatorios

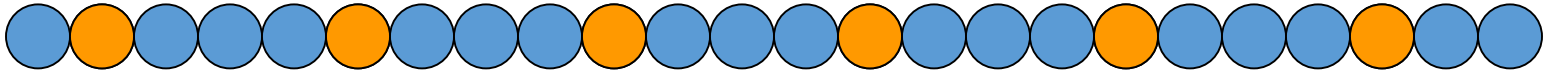
- RNG de Fibonacci Retardado (LFG):
  - $X_i = (X_{i-j} \star X_{i-k})\ mod\ M \quad 0 < j < k$
  - Donde $\star$ puede ser cualquier operador: suma, resta, multiplicación o XOR a nivel de bits.
  - Requiere inicializarse con *k* valores.
  - Los valores de j y de k se deben seleccionar cuidadosamente.

Generadores de Números Aleatorios

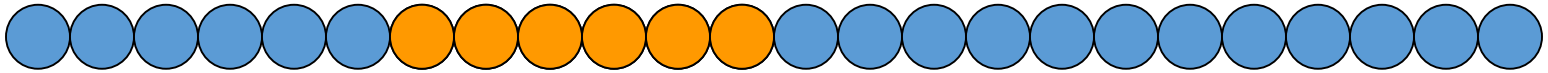
- Otras alternativas:
  - Generadores recursivos múltiples: Generalización de LFG.
  - Generadores no lineales:
    - Recursión cuadrática: $X_i = \left(a \times {X_{i-1}}^2 + b \times X_{i-1} + c\right) mod\ M$
    - Función inversa: $X_i = (a \times i + c)^{M-2}\ mod\ M$
  - Generadores combinados:
    - Barajado: Se usa un segundo generador para determinar el orden de los números generados.
    - Combinación de bits: Se combinan dos o mas generadores por medio de operadores (suma, resta u operadores de bits).

Generadores de Números Aleatorios en Entornos Paralelos

- Alternativas:
  - Generación desde un nodo único: No escala.
  - Generación de secuencias alternas: No es posible generar nuevos flujos aleatorios.



  - Particionamiento de secuencias: No reproducible y se pueden generar correlaciones de periodo largo.



  - Generación de secuencias independientes con diferentes semillas: Se puede aplicar LCG o LFG

Generación de Distribuciones No-Uniformes

- Los RNG generan distribuciones uniformes. ¿Cómo se hace un muestreo de una distribución Normal?:
  - Métodos basados la inversa de la función de distribución de probabilidad acumulada: $X \sim f(\cdot)$ entonces $F(X) \in [0,1]$ entonces si $Y \sim \mathcal{U}(0,1)$ podemos definir $X \sim F^{-1}(Y)$
  - Transformación de Box-Muller: Dadas $X, Y \sim \mathcal{U}(0,1)$

$$R = \sqrt{-2\ln(X)}$$
$$\Theta = 2\pi Y$$
$$Z = R\ \cos\Theta$$
$$W = R\ \sin\Theta$$

Implementación simplificada

```
do {
        a=(2*x)-1;
        b=(2*y)-1;
        r=a^2 * b^2
while (r<=0 || r>=1);
f=sqrt(-2*ln(r)/r);
z=f*a;
w=f*b;
```

Generadores de Números Aleatorios

- Implementaciones:
  - C/C++: random() mejor que rand().
  - MATLAB:
    - <v7.3: Algoritmo del Ziggurat (periodo ~2^{1492})
    - V7.4+: Mersenne twister (periodo ~2^{19937})
  - EXCEL: No muy bueno (periodo ~2^{23})
- Tests de aleatoriedad: DIEHARD.
- Generadores puramente aleatorios (www.random.org): ruido térmico de diodos Zener o ruido atmosférico.

Modelización de Sistemas Biológicos

Componente Aleatoria de los Modelos